Predicting Sign of Depression via Using Frozen **Pre-trained Models and Random Forest Classifier**

Hassan Alhuzali, Tianlin Zhang and Sophia Ananiadou

National Centre for Text Mining

Department of Computer Science, The University of Manchester, United Kingdom

Abstract

Predicting and understanding how various mental health conditions present online in textual social media data has become an increasingly popular task. The main aim of using this type of data lies in utilising its findings to prevent future harm as well as to provide help. In this paper, we describe our approach and findings in participating in sub-task 3 of the CLEF e-risk shared task. Our approach is based on pre-trained models plus a standard machine learning algorithm. More specifically, we utilise the pre-trained models to extract features for all user's posts and then feed them into a random forest classifier, achieving an average hit rate of 32.86%.

Keywords

Natural Language Processing, Mental Health, Social Media, Feature Extraction, Pre-trained Models

1. Introduction

There have been many previous iterations of the CLEF e-risk shared task over recent years [1], where the collective goal of these tasks is to connect mental health issues to language usage [2]. However, previous work in this area has not been able to produce convincing solutions that connect language to psychological disorders and it therefore remains a challenging task to produce accurate systems [2]. This year's CLEF e-risk-2021 shared task [3] provided three different tasks, which are focused on pathological gambling (T1), self-harm (T2) and depression (T3). In this paper, we only focus on T3 of the shared-task.

Depression is one of the most common mental disorders, affecting millions of people around the world [4]. The growing interest in building effective approaches to detect early sign of depression has been motivated by the proliferation of social media and online data, which have made it possible for people to communicate and share opinions on a variety of topics. In this respect, social media are an invaluable source of information, allowing us to analyse users who present sign of depression in real-time. Taking advantage of social media data in detecting early sign of depression helps to benefit individuals who may suffer it and their loved ones, as well as to give them access to professional assistance who could advocate their health and

D 0000-0002-0935-0774 (H. Alhuzali); 0000-0003-0843-1916 (T. Zhang); 0000-0002-4097-9191 (S. Ananiadou) © 0 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



CEUR Workshop Proceedings (CEUR-WS.org)

CLEF 2021 - Conference and Labs of the Evaluation Forum, September 21-24, 2021, Bucharest, Romania

Analyzali@postgrad.manchester.ac.uk (H. Alhuzali); tianlin.zhang@postgrad.manchester.ac.uk (T. Zhang); sophia.ananiadou@manchester.ac.uk (S. Ananiadou)

https://hasanhuz.github.io/ (H. Alhuzali); http://www.zhangtianlin.top/ (T. Zhang);

https://www.research.manchester.ac.uk/portal/sophia.ananiadou.html (S. Ananiadou)

well-being [5]. In the following sections, we describe our contribution to T3 of eRisk-2021, which focused on detecting early risk of depression from a thread of user posts on Reddit.

The rest of the paper is organised as follows: Section 2 provides a review of related work. Section 3 discusses some experimental details, including the data and task settings as well as evaluation metrics. Section 4 describes our method. Section 5 discusses our results while section 6 highlights negative experiments with Multi-task learning. We conclude in section 7.

2. Related Work

There is a large body of literature on early sign detection of depression [6, 7, 8, 9, 10, 11, 12, 13]. Some of these studies make use of the temporal aspect plus affective features in identifying early sign of depression. For example, Chen et al. [8] attempts to identify early sign of depression of Twitter users by incorporating a progression of emotion features over time, whereas Schwartz et al. [9] examined changes in degree of depression via Facebook users by taking advantage of sentiment and emotion lexicons. In addition, Aragón et al. [7] introduced a method called "Bag of Sub-Emotions (BoSE)" aiming at representing social media texts by using both an emotion lexical resource and sub-word embeddings. The choice of posted images and users' emotion, demographics and personality traits are also shown to be strong indicators of both depression and anxiety [6]. The above mentioned studies highlight the important role of both emotion features and the temporal aspect in early detection of depression on social media. Due to the increased interest in this area, the CLEF e-risk lab has run a sub-task of measuring the severity of depression since 2018.

Some of the participant teams in this shared task present different approaches, including those based on standard machine learning algorithms (ML), deep learning and transformer-based models. Oliveira [14] participated in the erisk shared-task of 2020 and proposed a model named "BioInfo". This model used a Support Vector Machine with different types of hand-crafted features (i.e. bag of words, TF-IDF, lexicons and behavioural patterns), and it ranked the top-1 model of the competition. Martinez-Castano et al. [15] was also one of the participant team who utilised BERT-based transformers, achieving competitive results to that of the BioInfo model.

Our work is motivated by research focused on ML algorithms [14] and transformer-based models [15]. Our work differs from these two studies in the following ways: i) Our method combines the two approaches instead of relying on one of them. In this respect, we use the former to learn a single representation per user while utilising the latter to train on the learned representations. ii) We use the "SpanEmo" encoder [16] that is trained on multi-label emotion dataset. iii) We do not fine-tune both the transformer-based models as well as SpanEmo encoder on the shared-task data. In other words, we only treat them as feature extraction modules.

3. Experiments

3.1. Data and Task Settings

For our participation in T3 of eRisk-2021 [17], we combine the 2019 and 2020 sets provided by the organisers, and then randomly sample 80% and 20% for training and validation, respectively.

Both sets consist of Reddit data posted by users who have answered the Beck's Depression Inventory (BDI) questionnaire [18]. The questionnaire contains 21 questions, and aims to assess the presence of feelings like sadness, pessimism, loss of energy, Self-Dislike, etc. To pre-process the data, we adopt the following steps. We firstly remove empty, duplicate and broken posts (i.e., those that either break the Reddit rule or are removed). Next, we tokenise the text, convert words to lower case, normalise URLs and repeated-characters. Table 1 presents the summary of all three sets, including the number of subjects/posts in the train, valid and test sets. The number of depression categories across the three sets is also included.

Table 1

Statistics of data.

	Train	Valid	TEST
#subjects	72	18	80
#posts	35,537	6,207	30,382
avg #posts/subject	493	344	379
#minimal subjects	11	3	6
#mild subjects	21	6	13
#moderate subjects	18	[p[4	27
#severe subjects	22	5	34

3.2. Evaluation Metrics

For evaluating the results of our submission, we used four metrics, which measure different proprieties (e.g. distance between correct and predicted answers). The four metrics are¹:

- Average Hit Rate (AHR): computes the percentage of predicted answers that are the same as the ground-truth responses.
- Average Closeness Rate (ACR): computes the absolute difference between predicted answers and the correct ones. In other words, CR measure evaluates the model's ability to answer each question independently.
- Average Difference between Overall Depression Levels (ADODL): computes the absolute difference between overall depression level score (i.e., sum of all the answers) and the real score.
- Depression Category Hit Rate (DCHR): evaluates the results among four categories which are based on the sum of all answers of the 21 questions. The four categories are minimal, mild, moderate and severe depression. DCHR computes the fraction of cases in which the produced category is equivalent to that of the real questionnaire.

Finally, the results of each metric are computed for each user, and are then averaged over all users in the data set.

¹Model details about the evaluation metrics can be found in Losada et al. [1]

4. Method

We develop a host of models based on neural networks for the task of predicting the severity of depression. In this work, we experiment with pre-trained language models (PLMs) both for feature extraction and for fine-tuning. For the former, the PLMs are used to extract a feature vector for each user, whereas they are fine-tuned directly on the depression data for the latter. Through extensive experiments, we observe that using PLMs for feature extraction achieves the best results on the validation set. We then train a random forest classifier on top of the extracted features to predict one of the possible answers for each question. Figure 1 presents an illustration of our framework.

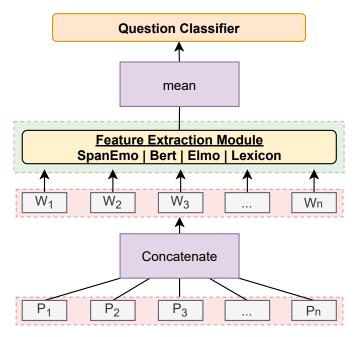


Figure 1: Illustration of our framework.

Let $\{p_i\}_{i=1}^N$ be a set of n posts, where each p_i consists of a sequence of M words = $(w_1, w_2, ..., w_M)$. As shown in Figure 1, the input to our framework is a list of all user's posts that are concatenated together. The output of this step, which is basically a sequence of words, is then fed into a feature extraction module (*f*). The feature extraction module computes the hidden representation for each user (*u*) as in equation (1):

$$\mathbf{u} = \frac{1}{M} \sum_{j=1}^{M} \mathbf{f}(w_j), \quad f(w_j) \in \mathbb{R}^d$$
(1)

where the above equation computes the mean over all tokens, with "d" denotes the dimensional size. This process attempts to obtain a single vector for each user that is ultimately fed to the classifier. Finally, each separate classifier is trained to predict one of the possible answers for each question. We now turn to describing the different types of feature extraction modules f used in this works as well as our implementation details.

Implementation Details. We used both PyTorch [19] and scikit-learn [20] for implementation and ran all experiments on an Nvidia GeForce GTX 1080 with 11 GB memory. Following the evaluation metrics discussed in 3.2, we run our experiments on average hit rate (AHR), average closeness rate, average of difference between overall depression levels (ADODL) and depression category hit rate (DCHR). In this work, we select three pre-trained models for feature extraction. Two of which are trained on a general domain (i.e., ELMo [21] and BERT [22]), whereas the third one (i.e., SpanEmo [16]) is trained on a similar domain to that of the shared-task. We briefly describe each of these models below:

- ELMo² is trained on a dataset of Wikipedia, which we use as our extraction module. More specifically, we extract the weighted sum of the 3 layers (word embedding, Bi-lstm1, and Bi-lstm2).
- Bert³ is trained on the BooksCorpus and Wikipedia. It includes a special classification token ([*CLS*]), which can be used as the aggregate input representation. The output of the ([*CLS*]) token is employed in this paper for feature extraction.
- SpanEmo⁴ is trained on the SemEval-2018 multi-label emotion classification data set [23]. It focuses on both learning emotion-specific features/associations and integrating the correlations between emotions into the loss function. We hypothesis that using a feature extraction model trained on a related domain to the problem under investigation can further boost the model performance compared to those models trained on a general domain.

5. Evaluation

Table 2 presents the results of our submission on all four metrics (i.e., AHR, ACR, ADODL and DCHR) and compares it to the top-ranked system on AHR. We submit three runs, where each one utilises different feature extraction module. For the first model, our feature extraction module is based on Elmo[21] plus some hand-crafted features (i.e., emotion dynamics [24] and Empath [25])⁵, whereas the second model makes use of Bert. Finally, the third model utilises SpanEmo-Encoder [16].

As shown in Table 2, the third model achieved the best results, thus demonstrating the utility and advantages of using a trained model on a related task to the one under investigation in this paper. This confirms our initial observation and helps to reinforce that our proposed model can benefit from the similarity between the two tasks in detecting sign of depression, given that some of the BDI questions are also related to emotion concepts, such as sadness, pessimism, loss of pleasure, self-dislike, etc.

Evaluating the Results of Different Layers. We also evaluate different layers in the SpanEmo-Encoder to determine which one is best for obtaining the highest results. The evaluation is presented in Figure 2, which reveals that different layers achieve different scores

²https://github.com/allenai/bilm-tf

³https://huggingface.co/transformers/index.html

⁴https://github.com/hasanhuz/SpanEmo

⁵We also added those features to both Bert and SpanEmo-Encoder, but the performance dropped, and hence we removed them.

Table 2

Model	AHR	ACR	ADODL	DCHR
RF (Elmo-Frzn+Feats)	31.43%	64.54%	74.98%	18.75%
RF (Bert-Frzn)	31.55%	65.00%	75.04%	21.25%
RF (SpanEmo-Encoder-Frzn)	32.86%	66.67%	76.23%	22.50%
DUTH_ATHENA (Top-AHR)	35.36%	67.18%	73.97%	15.00%

Experimental results on the test set. RF and Frzn: refers to the random forest classifier and the used weights of the respective feature extraction module, respectively.

depending on the chosen metric, especially for AHR and DCHR. Based on this evaluation, we selected the results of the ninth layer for our submission as it demonstrates strong performance on almost all four metrics.

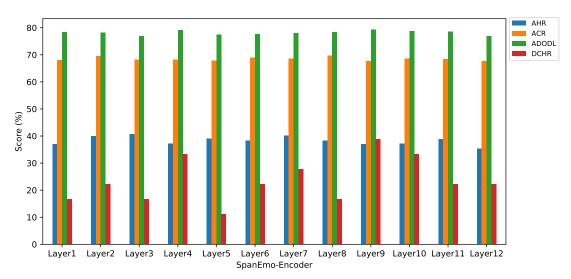


Figure 2: The results of each SpanEmo-Encoder layer when applied to the validation data set of the depression task.

6. Negative Results

We experimented with multi-task learning (MTL), for which we trained a single model for all 21 questions. More specifically, a shared BERT-based encoder was utilised to obtain a hidden representation for each post, and a specific head was then used for each question. To achieve a single output for each user, we aggregated the produced outputs from all posts via either averaging or summing. We also employed dynamic weighting of question-specific losses during the training process [26] as follows:

$$\mathscr{L}_{MTL} = \sum_{q}^{21} \frac{1}{2\sigma_q^2} \mathscr{L}_q + \log \sigma_q^2 \tag{2}$$

where q denotes a question and both \mathscr{L}_q and σ_q represent the question-specific loss and its variance. However, the results of MTL were not as high as the one discussed in section 4. This may be attributed to a number of factors. First, we used only a simple aggregate function that did not take the temporal aspect into consideration. This could be useful for detecting early sign of depression in users' posts⁶. Second, there was no annotations provided at the post-level which could help identify posts expressed severity sing of depression from those that do not. Third, we observed that the MTL model is overfitted with respect to the training data after the third or Fourth epoch although we used dropout to overcome that. This may be because the size of the data is quite small (i.e., roughly 70 users to train on), making the model unable to learn effectively from them.

7. Conclusion

We have proposed a framework aimed at detecting sign of depression from users' posts. We demonstrated that our proposed method achieved reasonable performance, especially for the AHR score. Our evaluation also showed that different SpanEmo-encoder layers produced different results. The choice of which layer to choose depends on the metric of interest. Finally, we reported some negative experiments, and hope that it will inspire the community to investigate further the vital role of learning a single model for all the 21 questions. This is motivated by the fact that some questions have some correlations/associations, and inferring the answer for one question may help infer others as well.

Acknowledgments

We would like to thank Saif Mohammad and Annika Schoene for fruitful discussions and their valuable feedback. The first author is supported by a doctoral fellowship from Umm Al-Qura University.

References

- D. E. Losada, F. Crestani, J. Parapar, Overview of erisk 2019 early risk prediction on the internet, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2019, pp. 340–357.
- [2] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk at clef 2020: Early risk prediction on the internet (extended overview) (2020).
- [3] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2021: Early risk prediction on the internet, in: In Proceedings of the Twelfth International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, Cham, 2021.
- [4] S. L. James, D. Abate, K. H. Abate, S. M. Abay, C. Abbafati, N. Abbasi, H. Abbastabar, F. Abd-Allah, J. Abdela, A. Abdelalim, et al., Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries

⁶Due to resource constraints, we could not train our model on user's timeline in a sequential minor.

and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017, The Lancet 392 (2018) 1789–1858.

- [5] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, J. C. Eichstaedt, Detecting depression and mental illness on social media: an integrative review, Current Opinion in Behavioral Sciences 18 (2017) 43–49.
- [6] S. C. Guntuku, D. Preotiuc-Pietro, J. C. Eichstaedt, L. H. Ungar, What twitter profile and posted images reveal about depression and anxiety, in: Proceedings of the international AAAI conference on web and social media, volume 13, 2019, pp. 236–246.
- [7] M. E. Aragón, A. P. López-Monroy, L. C. González-Gurrola, M. Montes, Detecting depression in social media using fine-grained emotions, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 1481–1486.
- [8] X. Chen, M. D. Sykora, T. W. Jackson, S. Elayan, What about mood swings: Identifying depression on twitter with temporal measures of emotions, in: Companion Proceedings of the The Web Conference 2018, 2018, pp. 1653–1660.
- [9] H. A. Schwartz, J. Eichstaedt, M. Kern, G. Park, M. Sap, D. Stillwell, M. Kosinski, L. Ungar, Towards assessing changes in degree of depression through facebook, in: Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality, 2014, pp. 118–125.
- [10] X. Wang, C. Zhang, Y. Ji, L. Sun, L. Wu, Z. Bao, A depression detection model based on sentiment analysis in micro-blog social network, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2013, pp. 201–213.
- [11] P. Van Rijen, D. Teodoro, N. Naderi, L. Mottin, J. Knafou, M. Jeffryes, P. Ruch, A data-driven approach for measuring the severity of the signs of depression using reddit posts., in: CLEF (Working Notes), 2019.
- [12] Y.-T. Wang, H.-H. Huang, H.-H. Chen, A neural network approach to early risk detection of depression and anorexia on social media text., in: CLEF (Working Notes), 2018.
- [13] F. Cacheda, D. F. Iglesias, F. J. Nóvoa, V. Carneiro, Analysis and experiments on early detection of depression., CLEF (Working Notes) 2125 (2018).
- [14] L. Oliveira, Bioinfo@ uavr at erisk 2020: on the use of psycholinguistics features and machine learning for the classification and quantification of mental diseases (2020).
- [15] R. Martinez-Castano, A. Htait, L. Azzopardi, Y. Moshfeghi, Early risk detection of self-harm and depression severity using bert-based transformers (2020).
- [16] H. Alhuzali, S. Ananiadou, Spanemo: Casting multi-label emotion classification as spanprediction, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021, pp. 1573–1584.
- [17] D. Losada, F. Crestani, A test collection for research on depression and language use, in: Proc. of Experimental IR Meets Multilinguality, Multimodality, and Interaction, 7th International Conference of the CLEF Association, CLEF 2016, Evora, Portugal, 2016, pp. 28–39.
- [18] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, J. Erbaugh, An inventory for measuring depression, Archives of general psychiatry 4 (1961) 561–571.
- [19] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch (2017).

- [20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
- [21] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proc. of NAACL, 2018.
- [22] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.
- [23] S. Mohammad, F. Bravo-Marquez, M. Salameh, S. Kiritchenko, Semeval-2018 task 1: Affect in tweets, in: Proceedings of the 12th international workshop on semantic evaluation, 2018, pp. 1–17.
- [24] W. E. Hipson, S. M. Mohammad, Emotion dynamics in movie dialogues, arXiv preprint arXiv:2103.01345 (2021).
- [25] E. Fast, B. Chen, M. S. Bernstein, Empath: Understanding topic signals in large-scale text, in: Proceedings of the 2016 CHI conference on human factors in computing systems, 2016, pp. 4647–4657.
- [26] A. Kendall, Y. Gal, R. Cipolla, Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7482–7491.